Αναφορά εργαστηριακής Άσκησης Εξόρυξη Δεδομένων και Αλγόριθμοι Μάθησης

2022-2023

Ανδρέας Τσιρώνης ΑΜ1063428

Όλες οι βιβλιοθήκες και τα εργαλεία που χρησιμοποιήθηκαν κατέβηκαν μέσω του Python package και εγκαταστάθηκαν μέσω του Python installer.

Η εργασία αναπτύχθηκε μέσω του jupyter notebook, χρησιμοποιώντας στην αρχή την εφαρμογή του jupyter μέσω browser και έπειτα μέσω extension του Visual studio code που επιτρέπει την εκτέλεση jupyter kernel.

Σαν βιβλιοθήκες χρησιμοποιήθηκαν οι εξής βιβλιοθήκες:

- Pandas, βασική βιβλιοθήκη για την διαχείριση των δεδομένων
- Numpy, επίσης βασική βιβλιοθήκη για την διαχείριση δεδομένων
- matplotlib (για την δημιουργία γραφικών παραστάσεων
- seaborn,χτισμένη πάνω στην matplotlib,ποιο όμορφες γραφικές παραστάσεις
- random, για επιλογή τυχαίας χώρας για γραφική αναπαράσταση.
- Sklearn, για την δημιουργία των clusters και του SVR
- Yellowbrick,για την δημιουργία των plot για το silhouette score (δεν λειτουργούσε τόσο καλά στο browser mode του jupyter notebook)

Όλα τα δεδομένα πάρθηκαν από την ιστοσελίδα που αναφέρεται στην εκφώνηση.

Περιγραφή υλοποίησης

Στο πρώτο μέρος της εργασίας, χωρίσαμε σε ξεχωριστά data frames τα σταθερά στοιχεία για κάθε χώρα και τα στοιχεία για κάθε μέρα σε ξεχωριστά dataframes. Δημιουργήσαμε καθημερινές τιμές για τα κρούσματα, τους θανάτους και την θετικότητα των τεστ. Υπολογίσαμε για κάθε χώρα τις μέσες τιμές των τεστ,κρουσμάτων,θανάτων και θετικότητας, όπως και την σχέση τους με τον πληθυσμό της κάθε χώρας. Για τον μέσο όρο των κρουσμάτων και των θανάτων, υπολογίσαμε μετά τα δέκα πρώτα κρούσματα, ώστε ο μέσος όρος να μην είναι πλασματικά μικρότερος από αυτό που είναι στην πραγματικότητα, καθώς έχει παρατηρηθεί ότι στις πρώτες μέρες, ή υπάρχουν τεστ χωρίς κρούσματα, ή κρούσματα όπου όμως περνάνε μέρες για να αυξηθούν, που δεν αντικατροπτίζει την πραγματική πορεία του ιού. Έτσι, κόβουμε πολλά από τις αρχικές τιμές που είχαν NaN πριν την συμπλήρωσή τους.

Σαν βασικές τιμές για να εξεταστούν για την επιτυχία ή όχι μίας χώρας ενάντια του ιού είναι η θετικότητα των test covid, η θνησιμότητα των ατόμων που έχουν τον ιό, στο σύνολο και σε μέρες που τα κρούσματα ξεπερνάνε το μέσο όρο, και το ποσοστό των κρουσμάτων και των θανάτων σε σχέση με τον πληθυσμό της χώρας. Αυτές τις τιμές χρησιμοποιούμε σαν βάση στα διαγράμματα και τις σχεδιάζουμε σε σχέση με τους άλλους συντελεστές τις κάθε χώρας

(gpd/capita,median age,θερμοκρασία κτλπ). Έπειτα εκτελούμε correlation matrix για να βρούμε τις διάφορες συσχετίσεις.

Στο δεύτερο μέρος, εκτελούμε k means clustering για να βρούμε ομάδες σε σχέση με τις πάνω 5 τιμές. Χρησιμοποιούμε και το elbow method και το SilhouetteVisualizer για το να βρεθεί ο καλύτερος αριθμός των clusters. Τα cluster είναι διαφορετικά σε νούμερο κάθε φορά, αλλά η διαμόρφωση των cluster είναι ίδια κάθε φορά. Χρησιμοποιούμε τις default τιμές για cluster,με n_init να είναι στο auto.

Μια σημαντική σημείωση που πρέπει να κάνουμε είναι ότι δεν μπορούμε να κάνουμε πραγματικές υποθέσεις πια χώρα πήγε καλά ή όχι η ίδια στην αντιμετώπιση του ίου, ή ήταν άλλοι παράγοντες που βοήθησαν στο να εμφανιστούν τα συγκεκριμένα νούμερα. Επίσης δεν μπορούν να κριθούν οι ποιότητες της αντιμετώπισης, όπως και η πραγματική επίδραση των κρουσμάτων και των θανάτων στο πληθυσμό.

Σχολιασμός των τελικών αποτελεσμάτων

Στους συσχετισμούς και στα διαγράμματα, αρχικά να αναφερθεί ότι φάνηκε ξεκάθαρα το γεγονός ότι συσχετίζεται υψηλή ηλικία, χαμηλή θερμοκρασία,gpd per capita, αριθμός των τεστ και αριθμός γιατρών είχε μία άμεση σχέση μεταξύ τους. Παρόλα αυτά, δεν είχε τόσο μεγάλη σχέση με τον αριθμό των κρεβατιών. Στις συσχετίσεις και στα διαγράμματα φάνηκε ότι η χαμηλή θετικότητα των τεστ (που συνήθως συνεπάγεται με πολλά τεστ),είχε πράγματι καλή επίδραση στο να πέσουν τα κρούσματα ή οι θάνατοι. Παρόλα αυτά, είδαμε ότι οι μετρικές που αναφέραμε (υψηλή ηλικία, χαμηλή θερμοκρασία,gpd per capita, αριθμός των τεστ και αριθμός γιατρών) είχε τάση να αυξήσει των αριθμών κρουσμάτων και θανάτων, κάτι που θεωρητικά θα περιμέναμε (χαμηλή θερμοκρασία,μεγάλη ηλικία),αλλά και όχι (πλούτος, αριθμός γιατρών,σε μικρό βαθμό αριθμός κρεβατιών).

Στους παρακάτω σχολιασμούς, θα πάρουμε τα στοιχεία που είδαμε στο αρχικό διάγραμμα που εμπεριέχει τις γενικές πληροφορίες για όλες τις χώρες και τις μέσες τιμές.

Η ομαδοποίηση των στοιχείων έδειξε μία ξεκάθαρη τάση 4 ομάδων, τα οποία φαίνεται να σχηματίζεται.

1. Ομάδα με το μεγαλύτερο αριθμό θνησιμότητας σε σχέση με τον πληθυσμό, με κάτω από τον μέσο όρο ποσοστό κρουσμάτων σχέση με τον πληθυσμό και το τρίτο μεγαλύτερο ποσοστό θανάτων σε σχέση με τον πληθυσμό. Επίσης έχουν υψηλότερο ποσοστό θετικότητας από το μέσο όρο και την μεγαλύτερη αύξηση της θνησιμότητας όταν τα καθημερινά κρούσματα πέρναγαν τον μέσο όρο.

   Αποτελείται κυρίως από χώρες που είναι στο μέσο όρο θερμοκρασίας, χαμηλό αριθμό κρεβατιών, γιατρών και πλούτου, υψηλότερο ποσό του 50% του ποσοστού του πληθυσμού από τον μέσο όρο και ίδια ηλικιακή κατανομή με τον γενικότερο πληθυσμό.

   Προσωπικός σχολιασμός: Αυτές οι χώρες δεν είχαν για ένα άγνωστο σύνολο παραγόντων, τόσο μεγάλη διασπορά ιού, άλλα δεν στάθηκαν τόσο ικανές να φροντίσουν τα άτομα που κόλλησαν τον ιό.

2. Ομάδα με το μεγαλύτερο αριθμό κρουσμάτων σε σχέση με τον πληθυσμό και το μεγαλύτερο ποσοστό θανάτων σε σχέση με τον πληθυσμό. Η θνησιμότητα, κανονική και εντατική, ήταν κοντά στο μέσο όρο και η θετικότητα των τεστ ήταν η καλύτερη.

   Η θερμοκρασίες έτειναν να είναι κάτω από το μέσο όρο, οι γιατροί και τα κρεβάτια πάνω από το μέσο, πλούτο ανά άτομο πάνω από το μέσο όρο και ηλικιακή κατανομή αρκετά πάνω από το μέσο όρο

   Προσωπικός σχολιασμός: Αυτές οι χώρες είχαν, για ένα άγνωστο σύνολο παραγόντων, μεγάλη διασπορά ιού, στάθηκαν αρκετά ικανές να φροντίσουν τα άτομα που κόλλησαν τον ιό, αλλά λόγω της μεγάλης διασποράς του ιού, οι συνέπειες του ιού ήταν μεγάλες στον πληθυσμό,

3. Ομάδα με το μικρότερο αριθμό κρουσμάτων σε σχέση με τον πληθυσμό και το μικρότερο ποσοστό θανάτων σε σχέση με τον πληθυσμό. Η θνησιμότητα, κανονική και εντατική, ήταν η μικρότερη και η θετικότητα των τεστ ήταν στο μέσο όρο, σε καλά επίπεδα.

   Το εύρος όλων των τιμών είχε την μεγαλύτερη διασπορά και τα περισσότερα outlier από κάθε ομάδα. Το κύριο ποσοστό κινήθηκε σε υψηλές θερμοκρασίες, σε χαμηλό αριθμό γιατρών και κρεβατιών (καλύτερο όμως από την πρώτη ομάδα), μικρό πλούτο ανά άτομο, και ηλικιακή κατανομή χαμηλότερη από το μέσο όρο.

   Προσωπικός σχολιασμός: Αυτές οι χώρες είχαν, για ένα άγνωστο σύνολο παραγόντων, τα καλύτερα αποτελέσματα. Ο ιός δεν διαδόθηκε τόσο, και αν διαδόθηκε, δεν είχε τόσους θανάτους, ακόμα και στις εντατικές μέρες.

4. Ομάδα με τον δεύτερο αριθμό κρουσμάτων σε σχέση με τον πληθυσμό και τον δεύτερο ποσοστό θανάτων σε σχέση με τον πληθυσμό, και τα δύο πάνω από τον μέσο όρο. Η θνησιμότητα, κανονική, ήταν επίσης σχετικά η και η θετικότητα των τεστ ήταν πάνω από τον μέσο όρο.

   Οι τιμές των γενικών τιμών κυμαίνονται στα πλαίσια της δεύτερης ομάδες αλλά με μεγαλύτερο εύρος και προς τα πάνω και προς τα κάτω.

   Προσωπικός σχολιασμός: Αυτή η ομάδα, για ένα άγνωστο σύνολο παραγόντων, κινήθηκε σε ένα ενδιάμεσο επίπεδο σε σχέση με τις άλλες τρεις ομάδες, τείνοντας περισσότερο όμως στην δεύτερη ομάδα. Οι τιμές στις δεν ήταν τόσο καλές όσο της τρίτης ομάδας, αλλά δεν μπορεί να κριθεί τόσο εύκολα όσο οι δύο πρώτες

Ο k means clustering και ο αριθμός των κέντρων στάθηκε πολύ χρήσιμος στο να αντλήσουμε στοιχεία την επιτυχία ή όχι των διαφόρων χωρών και τις σχέσεις μεταξύ τους, αλλά ίσως επιπλέον κέντρα να χρειαζόντουσαν για να διαχωρίσουν τις χώρες που υπάρχουν στην 3 και 4 ομάδα, όπου ήταν η μεγαλύτερες κατηγορίες και υπήρχαν οι πιο διαφορετικές μεταξύ τους χώρες .Αυτό όμως ελλοχεύει τον κίνδυνο να μπουν χώρες σε λάθος κατηγόριες, με βάση την ανάλυση που έδειξε το SilhouetteVisualizer